Proposal of a Brazilian Database Government Open Linked Data – DBgoldbr

Invited Paper

Marcio Victorino
University of Brasília
Information Science Faculty
Brasília, Brazil
5561999811923
mcvictorino@unb.br

Maristela Terto de Holanda
University of Brasília
Dept. of Computer Science
Brasília, Brazil
5561 993928784
mholanda@unb.br

Edison Ishikawa
University of Brasília
Dept. of Computer Science
Brasília, Brazil
5561 993928784
ishikawa@unb.br

Edgard Costa Oliveira
University of Brasília
Dept. of Production Engineering
Brasília, Brazil
5561992196093
ecosta@unb.br

George Ghinea
Brunel University London
Dept. of Computer Science UK
& Faculty of Technology, Westerdals
Oslo School of Arts, Communication
and Technology,Norway
44(0)1895266033
george.ghinea@brunel.ac.uk

Sammohan Chhetri
Brunel University London
Dept. of Computer Science UK
Uxbridge, Middlesex
44(0)1895266033
sammohan.chhetri@gmail.com

ABSTRACT

The Brazilian Government has made available on the Web a massive volume of public data. This data may be structured, semi-structured or non-structured in order to turn the administration as transparent as possible. Thus, we notice the great challenge in providing applications capable enough to handle this Big Data environment, and to make information available for decision making. In this environment, data processing is done via new approaches from Information Science and Computer Science areas, by involving Technologies and processes for collecting, representing, storing and disseminating information. This paper presents a conceptual model, the technical architecture and the prototype implementation of a tool DBgoldbr, designed to classify government public information with the help of ontologies, by transforming open data into open linked data. To fulfill the purpose of the solution, we used Soft System Methodology to identify problems, to collect users needs and to design solutions that fit the purpose of specific groups. The DBgoldbr tool was designed to ease up the search for open data made available by many Brazilian Government institutions, so that this data can be reused to support the evaluation and monitoring of social programs, in order to support the design and management of public policies.

MEDES'16, November 01-04, 2016, Biarritz, France
© 2016 ACM. ISBN 978-1-4503-4895-9/17/11…$15.00
DOI: http://dx.doi.org/10.1145/3012071.3012093

CCS Concepts

●Information Systems → Information retrieval → Search engine architectures and scalability → Search engine indexing

Keywords

Open data, linked data, open government, Brazilian open data infrastructure.

INTRODUCTION

The Brazilian government has made available a massive volume of public data in order to pursue the principle of transparency in the general administration. This data is available in a structured, semi-structured or not structured at all. It is available on line (Transparency Portal[1] of Brazilian Government) access to data on public expenses of some of its representatives, public institutions such as INEP[2] or IBGE[3], huge Brazilian data management public institutions responsible for all national demographic data and educational data, for example, and many others.

[1] http://www.transparencia.gov.br/
[2] http://portal.inep.gov.br/inep-data
[3] http://www.ibge.gov.br/

In this context rises the need to develop integrated applications that are able to generate insights in an appropriate speed out of a huge number of data in varied formats, the so-called Big Data Applications Environment. Processing this information resource demands new research from computer science and information science, in all phases of the information cycle.

Due to its complexity, we foresee the future implementation of a Big Data Ecosystem that can support the analysis of linked open data from the Brazilian government. Via a deep study on this matter, we verified that the ecosystem proposed by [1] is aligned to the government goals as well. We also consider the use of ontologies to support semantic interoperability.

The proposed big data ecosystem will allow data storage from different sources to be treated and to support evaluation of social programs and to support public policies management. The extension of this ecosystem consists of our proposed tool, built with a set of components that use an ontology-based semantic classification of information, called DBgoldbr. Its main goal is to transform open data into linked open data and to ease up the identification and localization of these data sources. This is done from the semantic description or from its metadata. This paper presents our proposal of a conceptual view as well as a technical architecture of the DBgoldbr.

GOVERNMENT LINKED OPEN DATA

According to the Open Knowledge Foundation , "open means anyone can freely access, use, modify, and share for any purpose (subject, at most, to requirements that preserve provenance and openness)". The World Wide Web Consortium – W3C defines open government data as the publication and disclosure of public information on the web, shared in an open gross format, logically comprehensible allowing the reuse of information in digital applications developed by society. Eaves is an open data activist and considered one of the main experts on the matter, and also a specialist in public policies. He proposed three laws that were also adopted by the W3C [2]: If it can't be found or indexed, it doesn't exist; If it isn't available in open and machine readable format, it can't engage; If a legal framework doesn't allow it to be repurposed, it doesn't empower.

BIG DATA AND SEMANTIC CLASSIFICATION OF INFORMATION

There are various definitions and understandings for the term "Big Data". One of the most widely accepted is the 3V definition presented by Laney [3] and ratified by Beyer and Laney [4]: "Big data is high-volume, high-velocity and high-variety information assets that demand cost-effective, innovative forms of information processing for enhanced insight and decision making.". However, Demchenko, Y. et al. in [5] proposed the Big Data definition as having the following 5V properties: Volume, Velocity, Variety that constitute native/original Big Data properties, and Value and Veracity as acquired as a result of data initial classification and processing in the context of a specific process or model. Uddin and Gupta in [6] proposed 7V thus including volatility and visualization. [6] Volatility refers to data meaning, which is constantly changing and visualization relates to all means and that is easy to understand and read.

There are some definitions of Big Data Ecosystems, such as Shin and Choi [7] for instance, where big data is seen as an ecological ecosystem that involves the following aspects: technology, government, industry, market, users and society, taking into account the effects of big data in all involved sectors.

Big Data is defined by [1] as a complex system of technical facilities and components built around a source of data specific and its application: the complex of interrelated components is used to storage, processing, visualization and delivery of the results obtained from the big data. This ecosystem comprehends the big data itself as well as the following categories of architectural components:

1) Models and data structures: according to [1], the many stages of big data transformation require different data structures, models and formats, including the possibility to process both structured and non-structured data. It is possible that the data structure and corresponding models suffer changes during the different stages of data processing. However, it is important to keep the link between these structures.

Figure 1 shows examples of structures, models and links of data, from the original figures from [1]. The top of the figure represents a data model containing information such as: structures and data types, links between data: raw data into information, and information to presentation. Raw data represents data in its original state, as it was brought from its original source. Below they are represented with arrows the origins for data transformations during the processing cycle.

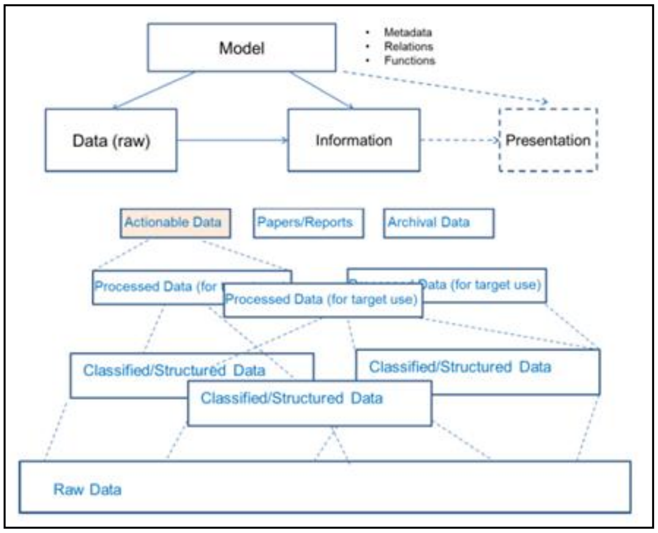

Figure 1 – Big Data structures, models and their links at different processing stages.

2) Big Data Architecture: it is formed by a set of Technologies and componets for the Big Data processing and analysis. In [1] Demchenko et al. mention two groups of main Technologies named Big Data Analytics Infrastructure – BDAI, which are: a general architecture formed by technologies and

components for storage, computing, network, devices and operational support to Big Data; and the processing and analysis architecture, which is formed by tools for data presentation and visualizations, as well as analysis and processing.

3) Big Data Life Cycle Management: [1] emphasize the need to use scientific methods to obtain the benefits of new opportunities to data collection and mining in order to acquire the necessary information. The information lifecycle involves storage and preservation of data in different stages, in order to allow the reuse of analytic research of processed data and published results. However, this is only possible if necessary solutions has been implemented to allow complete identification, crossed referencing and data linkage. Data integrity, control and auditing must be supported during all data lifecycle.

Figure 2 shows the top storage layer, where data is persisted and data models that represent them during the whole life cycle. In the bottom part of the figure, we present the phases of data transformation, starting with data collection (and data registration), data treatment (filtering, enriching and classification of data) processing and analysis of data (analytics, modelling and prediction of data), delivery of results generated to the consumer data analytics application, via the delivery and visualization of data process.

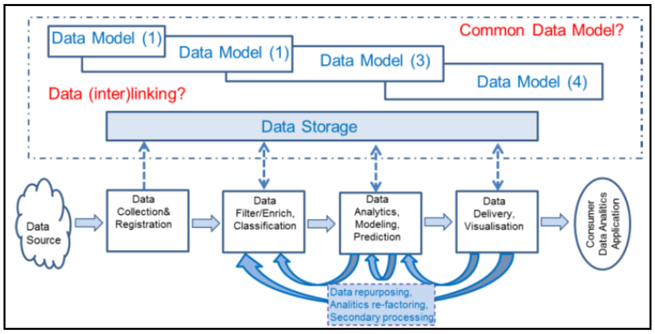
Data Model (1)
Data Model (1)
Data Model (3)
Data Model (4)
Common Data Model?
Data (inter)linking?
Data Storage
Data Source
Data Collection& Registration
Data Filter/Enrich, Classification
Data Analytics, Modeling, Prediction
Data Delivery, Visualisation
Consumer Data Analitics Application
Data repurposing, Analitics re-factoring, Secondary processing

Figure 2 – Big Data Lifecycle in Big Data Ecosystem.

4) Big Data security infrastructure: it gathers the necessary set of components and policies to provide data access control and a safe processing environment.

The ecosystem to treat open data in the Brazilian government is planned within the extension of the Big Data Ecosystem from [1], by introducing and additional structure for semantic classification of sources in the category of components to manage big data lifecycle, and thus providing data linkage via its semantic representation.

This semantic classification structure is what we propose as the Brazilian Database Government Open Linked Data - DBgoldbr, which is being implemented by using the W3C standards, via the resource description framework - RDF, SPARQL Protocol and RDF Query Language - SPARQL, Uniform Resource Locator – URL e Web Ontology Language – OWL. This allows information sources to reach a level 4 star for data openness. DBgoldbr can also create references to other open data sources of the government, because it is structured to semantically connect open data as they are made available, with the help of an ontology of the government state that allows a semantic linkage between government sources, by then providing a 5 star open data level context.

Based on the relevance of interoperability of open government data, Berners-Lee in [8] proposed the 5 star system, which classifies the degree of data openness, the more open the higher the stars and the easier it is to enrich the data.

The 5 stars of open linked data are:

- "1 star": make your stuff available on the Web (whatever format) under an open license;
- "2 stars": make it available as structured data (e.g., Excel instead of image scan of a table);
- "3 stars": make it available in a non-proprietary open format (e.g., CSV instead of Excel));
- "4 stars": make it available in a non-proprietary open format (e.g., CSV instead of Excel) (Resource Description Framework - RDF e SPARQL Protocol and RDF Query Language - SPARQL): use of Uniform Resource Locator – URL;
- "5 stars": link your data to other data to provide context."

An analysis made with the available Brazilian Government data from the public administration entities showed that a big amount of the data presents level 3 stars, because they are available on the Internet in a structured way, in a non-proprietary format CSV, and any person or entity can make downloads. However, we hope to raise the openness level of data from level 3 to level 5, by making use of W3C standards and ontologies to transform open data into open linked data.

DATABASE GOVERNMENT OPEN LINKED DATA – DBgoldbr CONCEPTUAL MODEL

DBgoldbr aims to do a semantic classification of already published data sources. This semantic classification is based on the use of ontologies and RDF triples in order to transform open published data into open linked data. Figure 3 presents a conceptual view of how these resources are used.

As we can see on figure 3, all data sources are published in the CSV format and are represented via URIs and are part of the triples as subjects. The data sources, located in the government institutions web sites, will be part of the triples as objects. The predicates are based on controlled vocabularies and for that purpose we use ontologies. The generated triples can be stored in RDF XML, N-Triples, Turtle or JSON. Therefore, we expect a big volume of RDF triples to represent information, the environment uses an RDF Database Management System to store the triples, which is a persistency device available from the Jena Apache Project [9].

By using DBgoldbr, we hope to enhance the quality of queries search results that users made in the open data sources from the Brazilian government on the web.

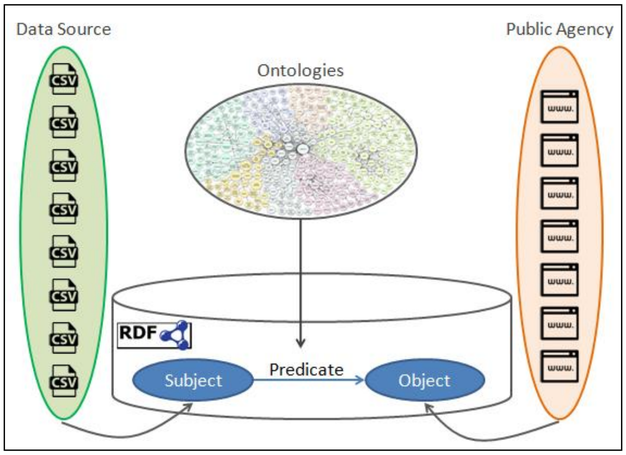

Figure 3 – Conceptual View of Database Government Open Linked Data - DBgoldbr.

Currently one of the most popular search tool used by the users is Google's search engine. Figure 4 presents a simplified representation of the conventional search process by a user searching for open data sources published by the Brazilian government.

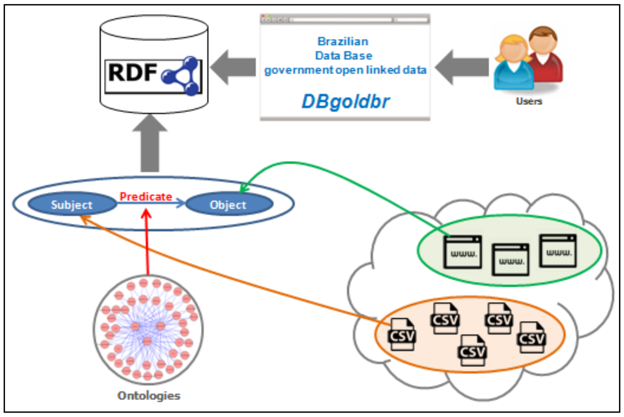

Figure 4 – Using DBgoldbr to search for open data sources.

DBgoldbr aims at replacing these conventional search tools, during the search, by open data sources made available by the Brazilian Government. We built an RDF triple repository with a semantic description of the data sources. Initially, these triples are created instantly as the source is made available and then they use one of the DBgoldbr interfaces. There is a need to define an automated classification technique in order to perform a data mining. Figure 4 presents the source search process using DBgoldbr.

Figure 4 shows DBgoldbr with a native repository of RDF triples that semantically describe the published sources (subject), the publishing entities (object) and the predicates which are the terms obtained from the ontologies. This feature offers users with search options based on the available data sources, obtained from the institutions that have published data or vocabularies. In turn, this offers a common understanding within the domain and thus improve the quality of results. The DBgoldbr resources are presented in details in the following sections.

LOGICAL MODEL OF THE DBgoldbr

The logical architecture of the DBgoldbr is organized in two parts: the first part is composed of the registration of publishing entities, published sources and ontologies; the second part involves the necessary resources for the creation, storage and query of RDF triples, described as follows.

First part of the DBgoldbr architecture

Figure 5 shows the first part of the logical architecture of the DBgoldbr, represented in layers: 1st layer is the application developed in Java (jdk1.8.0_101), the second layer is the Application Program Interface (API) Java Database Connectivity (JDBC), the third layer is the connection driver of the Database Management System (DBMS) MySQL and the forth layer is the DBMS MySQL (version 5.6.22.0). The second and third layers are used to allow the connection between the application and the database, based on Oracle and MySQL, respectively.

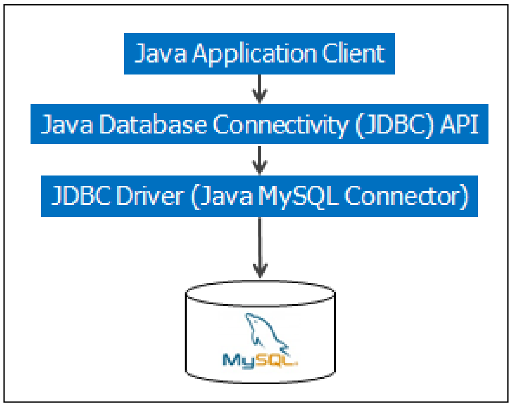
Java Application Client
Java Database Connectivity (JDBC) API
JDBC Driver (Java MySQL Connector)
MySQL

Figure 5 – First part of the Logical Architecture of DBgoldbr.

In this architecture, the Relational Database MySQL is the persistence device, to be used to persist data over the publishing entities (government institutions or any other that is capable of generation open data for the government). The database also persist the published sources and the ontologies in relational tables. In this figure, the Java Application Client layer is developed via the implementation of a prototype of DBgoldbr in Java.

Second part of the Logic Architecture of the DBgoldbr

Figure 6 presents the second part of the logic architecture of DBgoldbr, represented in layers, as follows: first layer a Java Application Client, representing the application, which is being developed in Java (jdk1.8.0_101), and the other layers are part of the Apache Jena framework. This framework has APIS to process

RDF files, which represent the triples, the OWL files, which represent the ontologies and give support to SPARQL queries in the RDF triples. There is also an API to allow inferences and triple stores. DBgoldbr uses TDB to persist RDF triples.

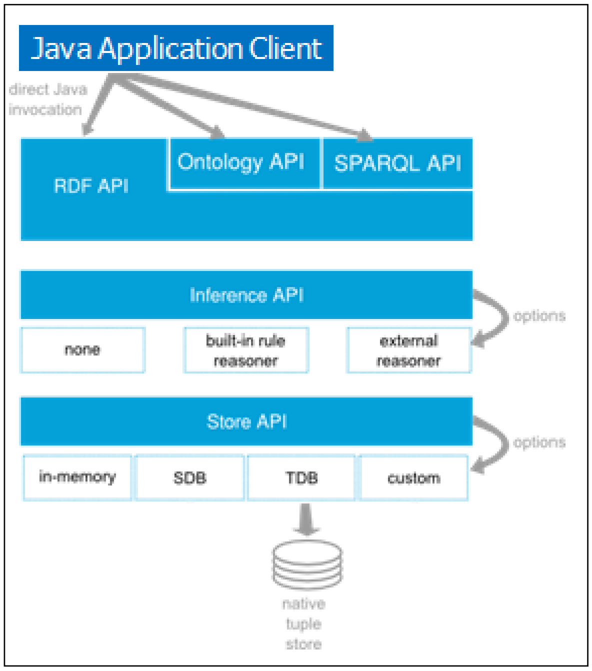

Figure 6 – Second part of the logical architecture of DBgoldbr.

CONCLUSIONS

This paper presents a solution to the data government of Brazil have disclosed in an open & linked manner, it has been designed for the benefits of citizens and institutions taking consideration of all the best practices. We proposed here a solution to this in the context of a Brazilian government open data disclosure to open access via Web.

However, we also learned that it is not simple to process massive volume of a daily generated by the Brazilian government institutions, in an environment of knowledge sharing with different formats and a varied IT infrastructure. This is a key process to enhance public efficiency and transparency as well as to allow public managers and staff to take daily safe decisions.

This paper presented the conceptual view and technical architecture of the Big Data Ecosystem DBgoldbr - Brazilian Database Government Open Linked Data – by illustrating the development of a prototype tool built with a set of resources to perform ontology-based semantic classification of information.

DBgoldbr aims at making available data compatible with the 5 stars level of open government data that can be connected via ontologies, so that we may organize and represent huge volumes of massive data and its respective semantics.

The following steps in this research is a plan made by the team of researchers from Computer Science, Information Science to migrate the software architecture to a web environment. We intend to offer the user end of DBgoldbr queries to a huge volume of public data in the many different areas of the government. The information professional may also identify relevant sources of information to prepare the appropriate decision making environment based on semantically represented linked open data mining. This solution is also designed to the public searching for data that can be integrated in order to help answer questions about the efficiency of public policies in social contexts, improving resource management and rationality of public expenditure. Citizens in general are mostly benefited by this solution due to the fact that public information is accessible to all users in an interoperable manner and intuitively (semantically) available, via open data repository interfaces as easy to use and integrate as the solution DBgoldbr described here.

REFERENCES

[1] Demchenko, Y. et al. 2014. Defining architecture components of the Big Data Ecosystem. In Collaboration Technologies and Systems (CTS), 2014 International Conference on (pp. 104-112). IEEE.

[2] Eaves, D. 2009. The three laws of open government data. Eaves. ca, 30.

[3] Laney, D. 2001. Application delivery strategies. META Group, Stamford. Also available at < http://blogs.gartner.com/doug-laney/files/2012/01/ad949-3D-Data-Management-Controlling-Data-Volume-Velocity-and-Variety.pdf>

[4] Beyer, M. A. and Laney, D. 2012. The importance of 'big data': a definition. Stamford, CT: Gartner, 2014-2018.

[5] Demchenko, Y. et al. 2013. Addressing big data issues in scientific data infrastructure. In Collaboration Technologies and Systems (CTS), 2013 International Conference on (pp. 48-55). IEEE.

[6] Uddin, M. F. and Gupta, N. 2014. Seven V's of Big Data understanding Big Data to extract value. In American Society for Engineering Education (ASEE Zone 1), 2014 Zone 1 Conference of the (pp. 1-5). IEEE.

[7] Shin, D.H. and Choi, M.J. 2015. Ecological views of big data: Perspectives and issues. Telematics and Informatics, 32(2), pp.311-320.

[8] Berners-Lee, T. 2006. Linked data-design issues. Also available at <http://www.w3.org/DesignIssues/LinkedData.html>

[9] Jena, A. 2013. Apache jena. Also available at <http://jena. apache.org>